SIGGRAPH ASIA 2016 MACAO

Sponsored by

3D Attention-Driven Depth Acquisition for Object Identification

Kai Xu, Yifei Shi, Lintao Zheng, Junyu Zhang, Min Liu, Hui Huang, Hao Su, Daniel Cohen-Or and Baoquan Chen

National University of Defense Technology   Shandong University

Shenzhen University   SIAT   Stanford University   Tel-Aviv University

SA2016.SIGGRAPH.ORG

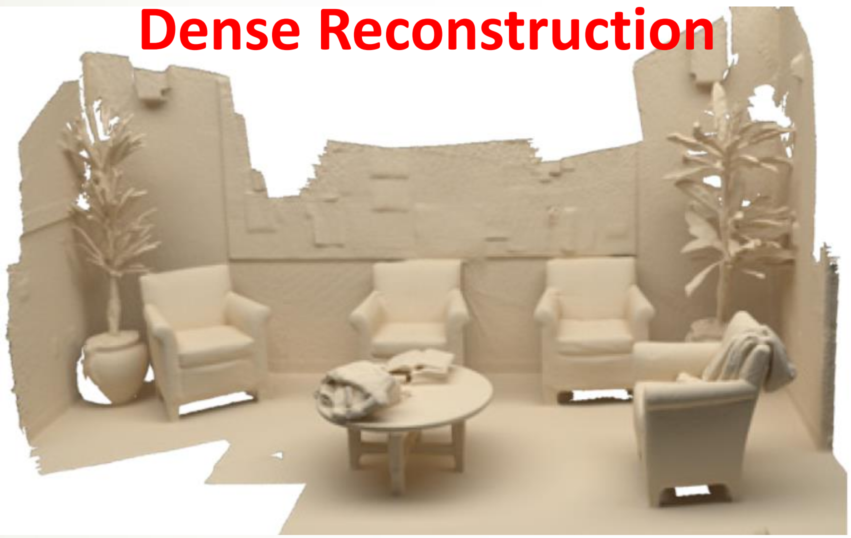
Background & motivation

- Robotic indoor scene modeling

Perception on object

Background & motivation

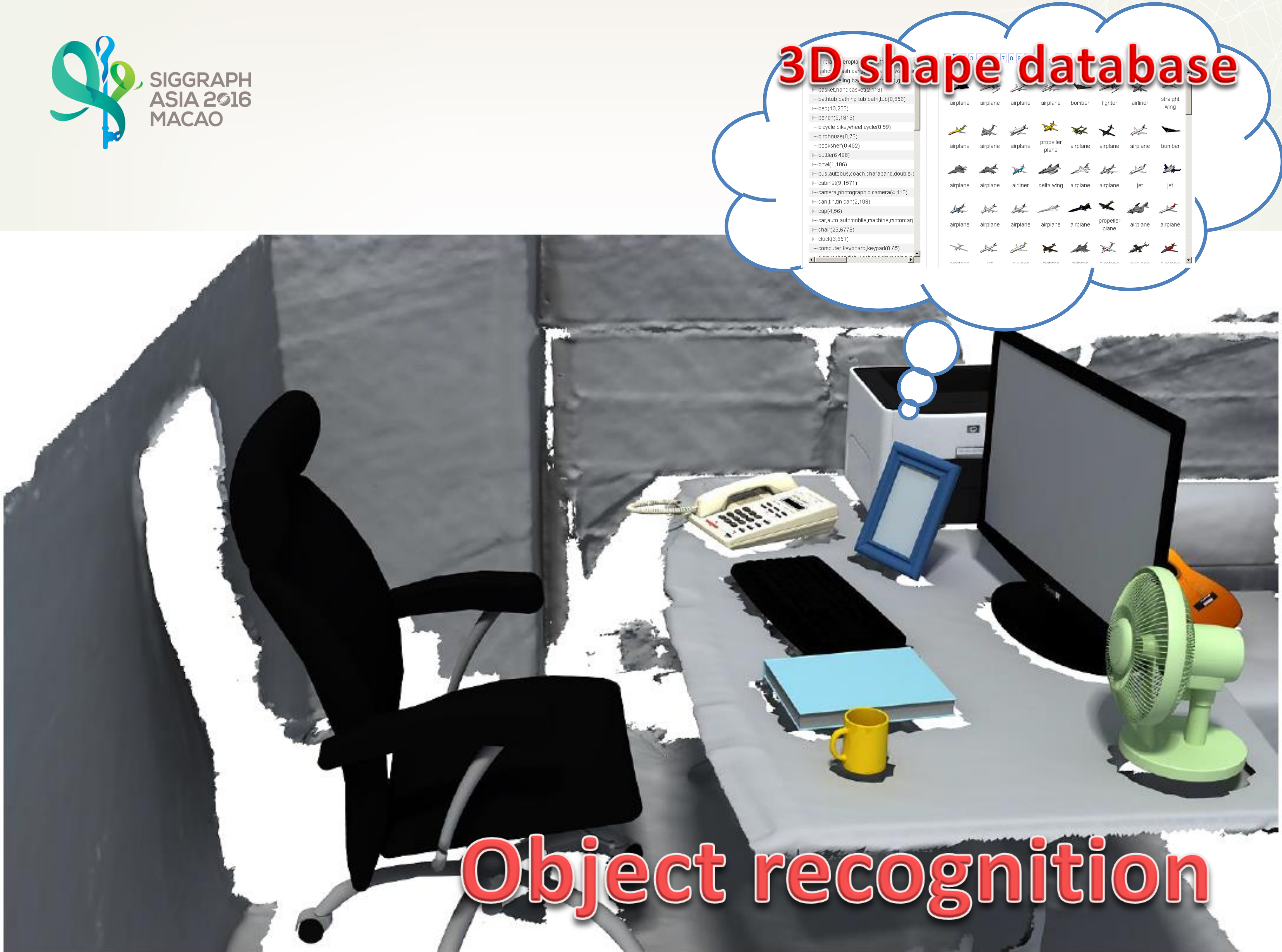
- Indoor environments acquisition and modeling



[Nießner et al. 2013]

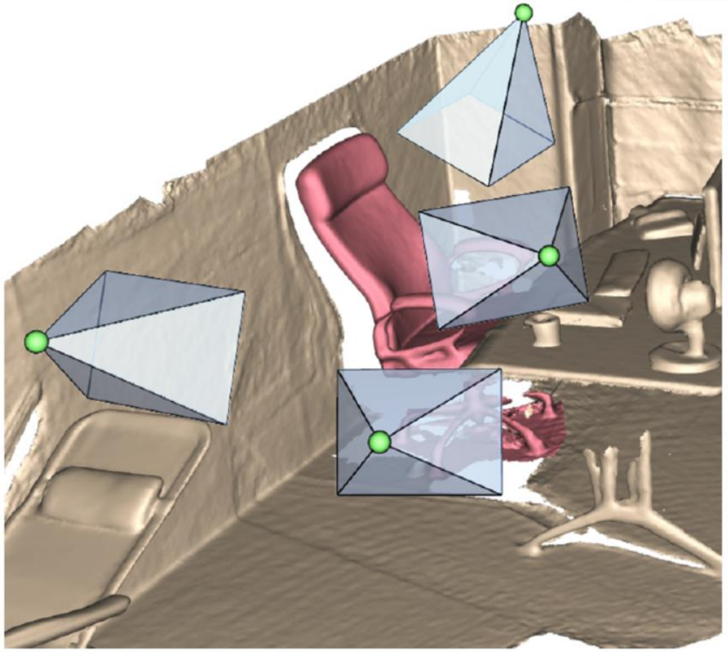
[Xu et al. 2015]

Background & motivation


3D shape database
Object recognition

SIGGRAPH ASIA 2016 MACAO

Active object recognition



Active object recognition

Problem setting

- A robot actively acquires new observations to gradually increase the confidence of object recognition
- Two key components:

Object classification

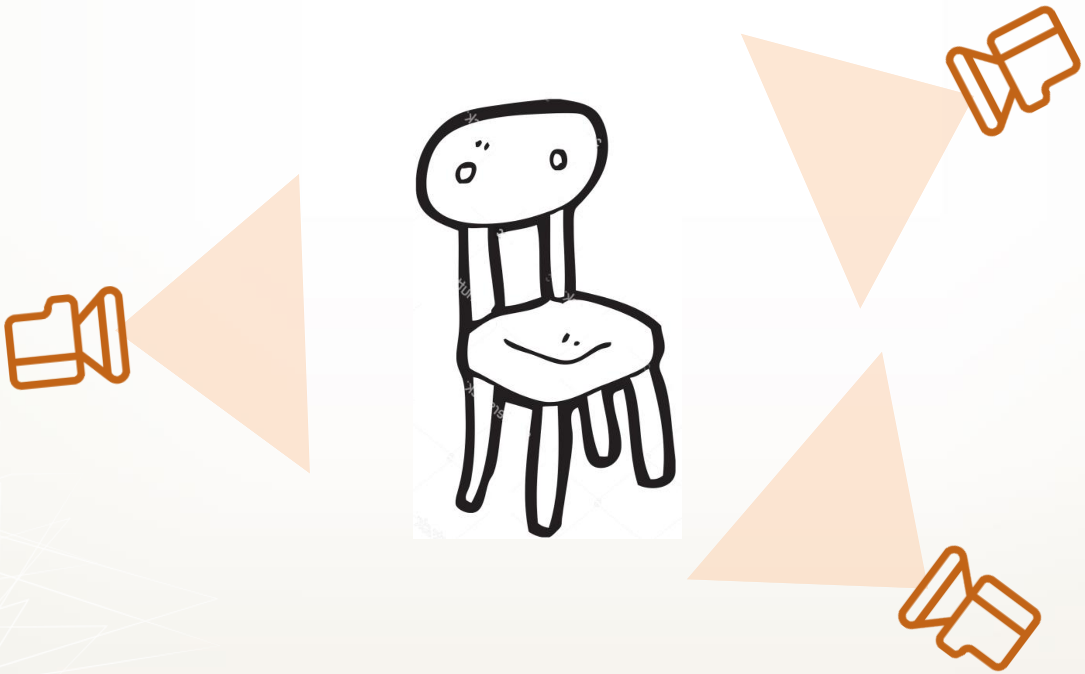
Estimate object class based on so far acquired observations

View planning

Predict the Next-Best-View to maximize its information gain

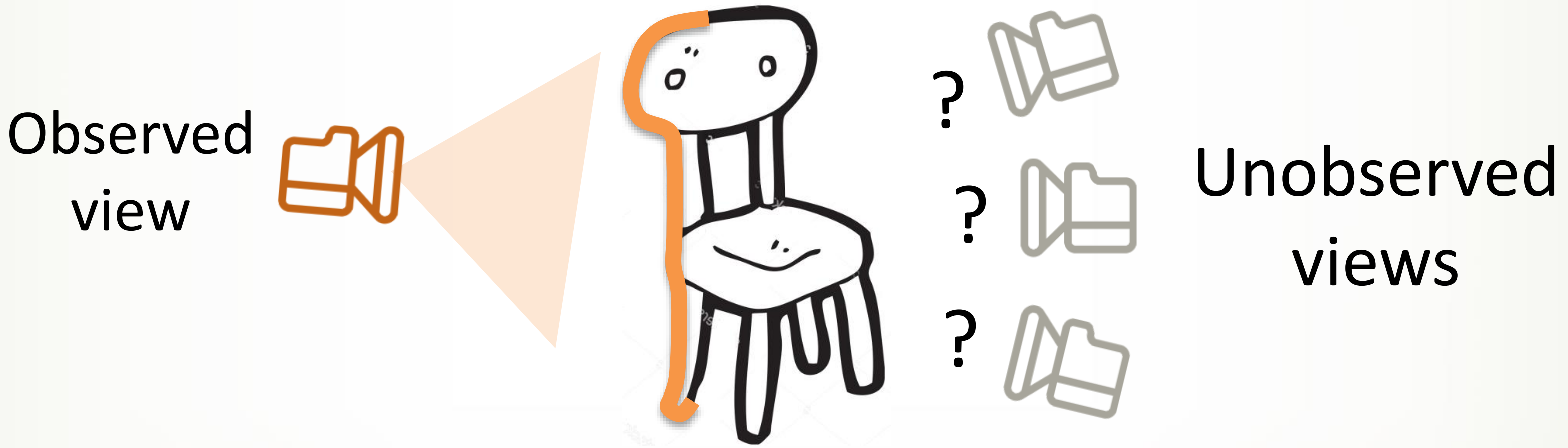
The main challenge

- **Observation is partial and progressive**
  - Shape description/matching with partial data is hard
  - Observations from varying views



The main challenge

- Observation is partial and progressive
  - View planning



How can you know which view is better without knowing its observation?

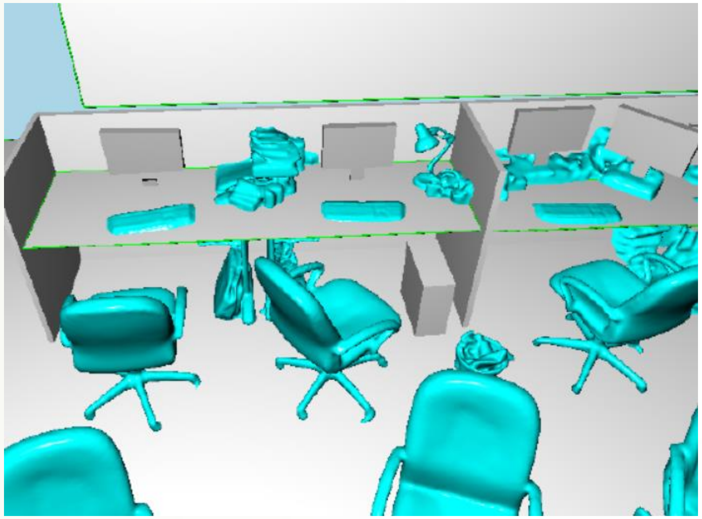
The main challenge

- Real indoor scenes are often cluttered
  - Degrade recognition accuracy
  - Invalidate the off-line learned viewing policy

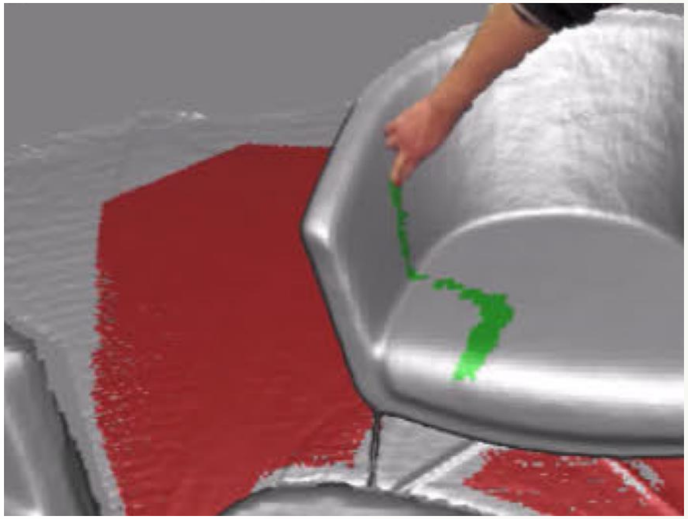
Related work

Related work

- Online scene analysis and modeling



Plane/Object Extraction
[Zhang et al. 2014]

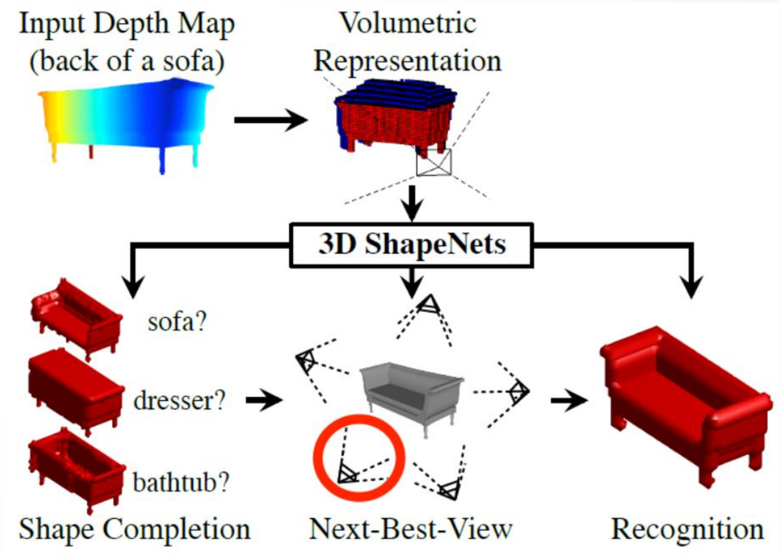


SemanticPaint
[Valentin et al. 2015]

Related work

- Active reconstruction and recognition

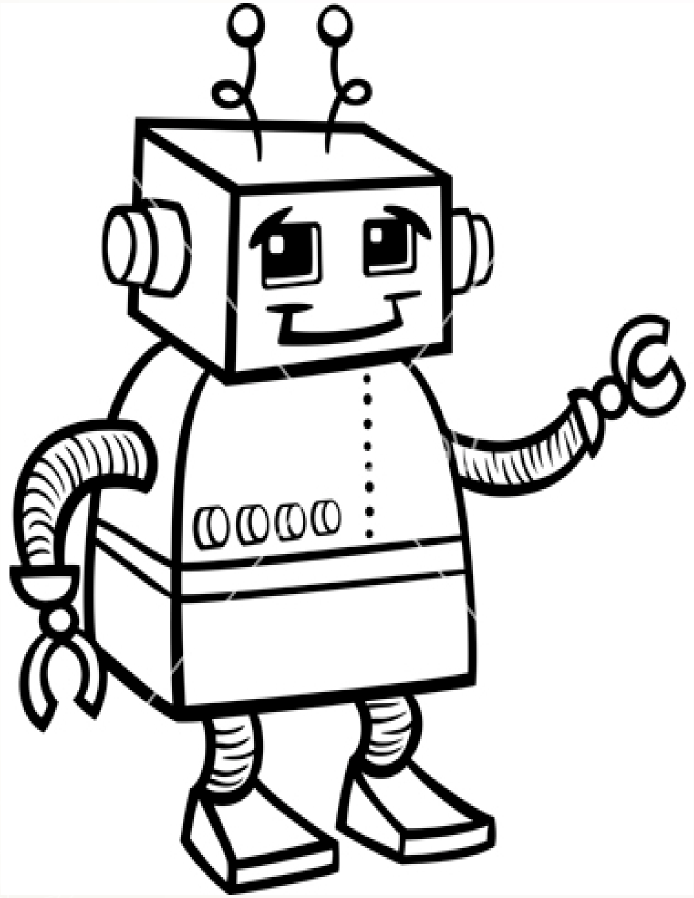
Next-best-view for reconstruction [Wu et al. 2014]



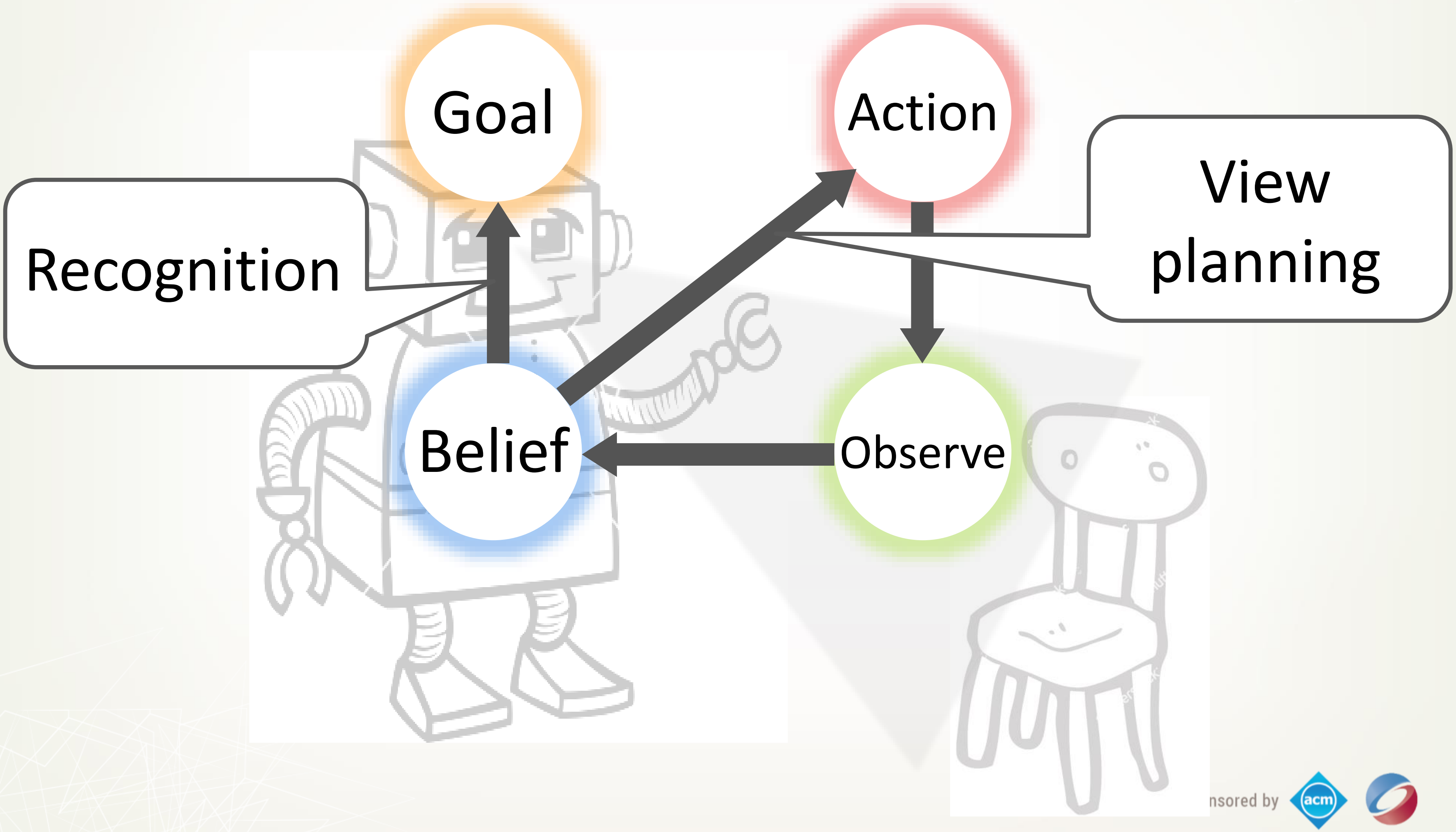
Next-best-view for recognition [Wu et al. 2015]

Method

The general framework



The general framework

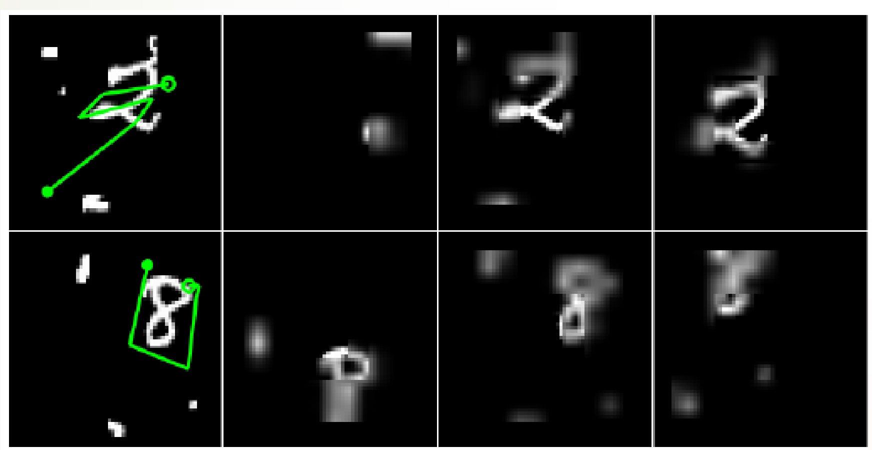


An attentional formulation

"Humans *focus attention selectively on parts* of the visual space to acquire information when and where it is needed, and combine information from different fixations over time to build up an *internal representation* of the scene"

— Ronald Rensink

Internal representation

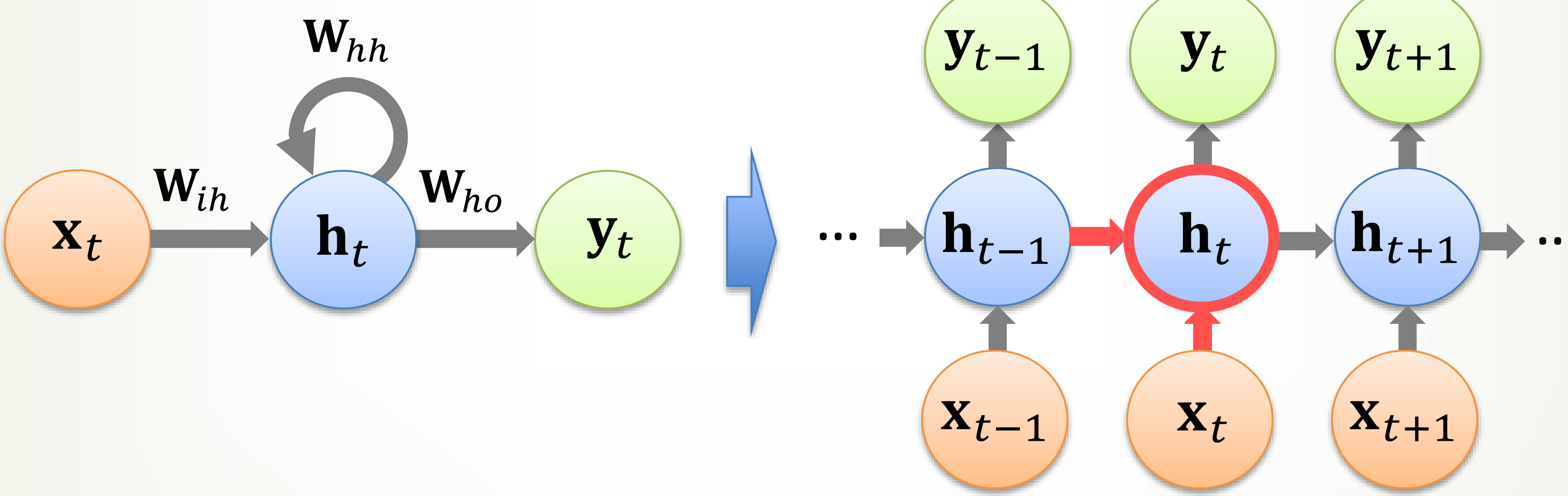


Hand-writing recognition
[Mnih et al. 2014]

Image caption generation
[Xu et al. 2015]

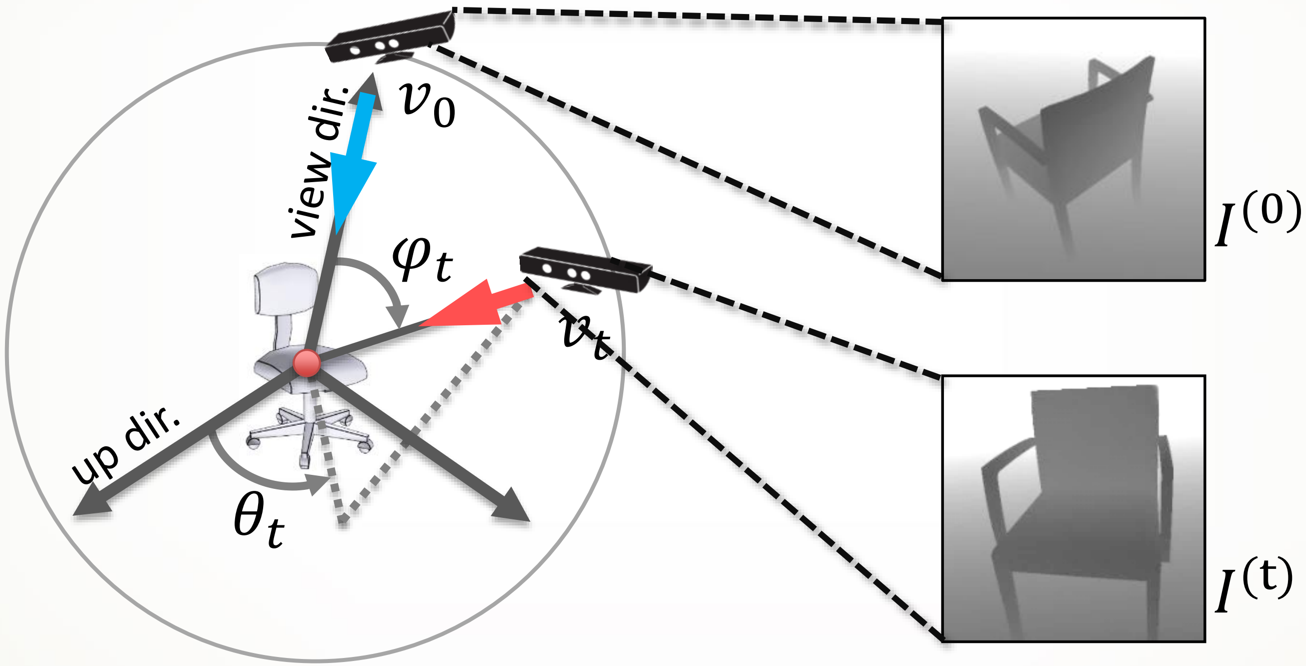
Recurrent Attention Model

- Recurrent Neural Networks (RNN)

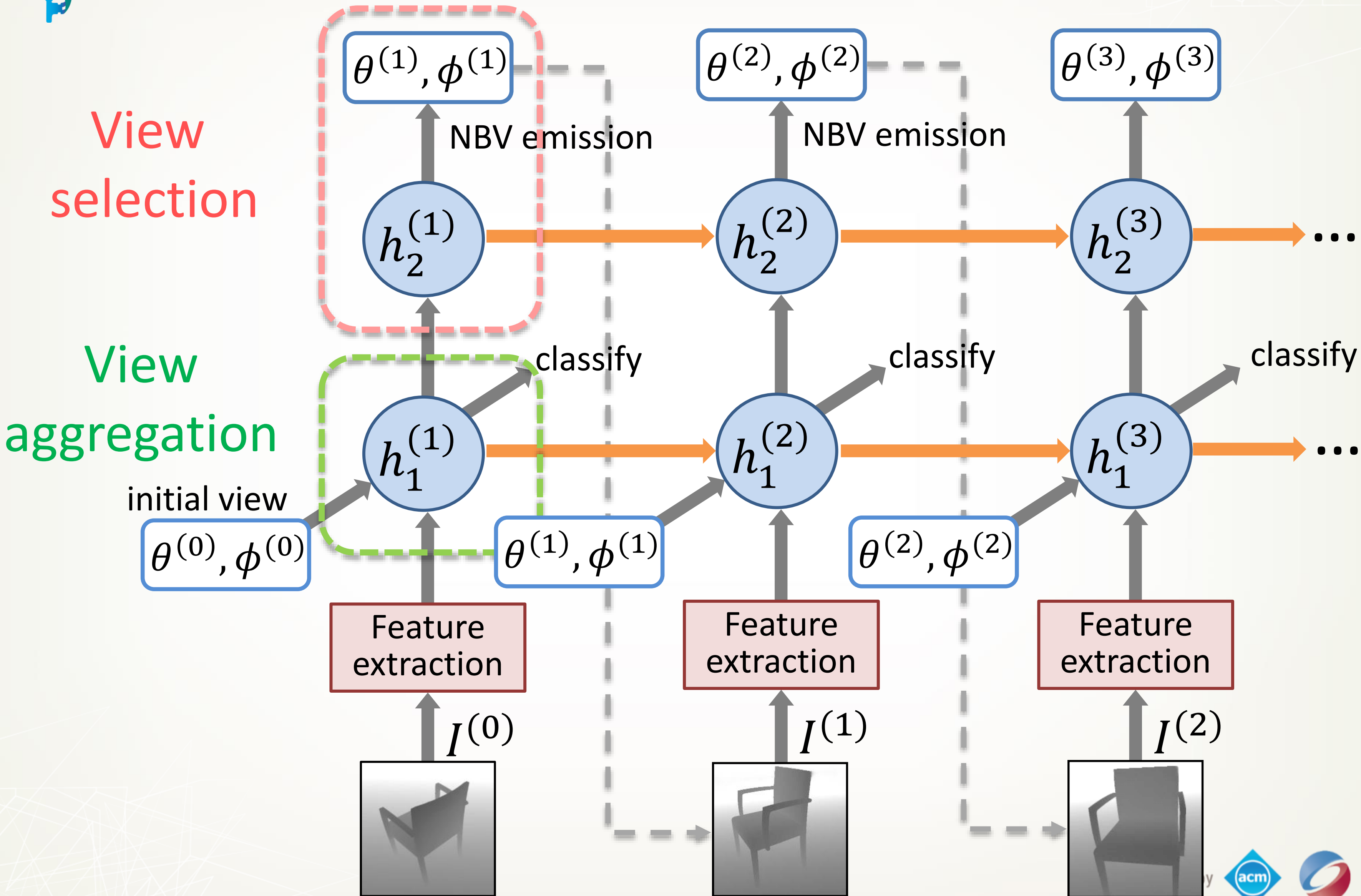


Aggregate information

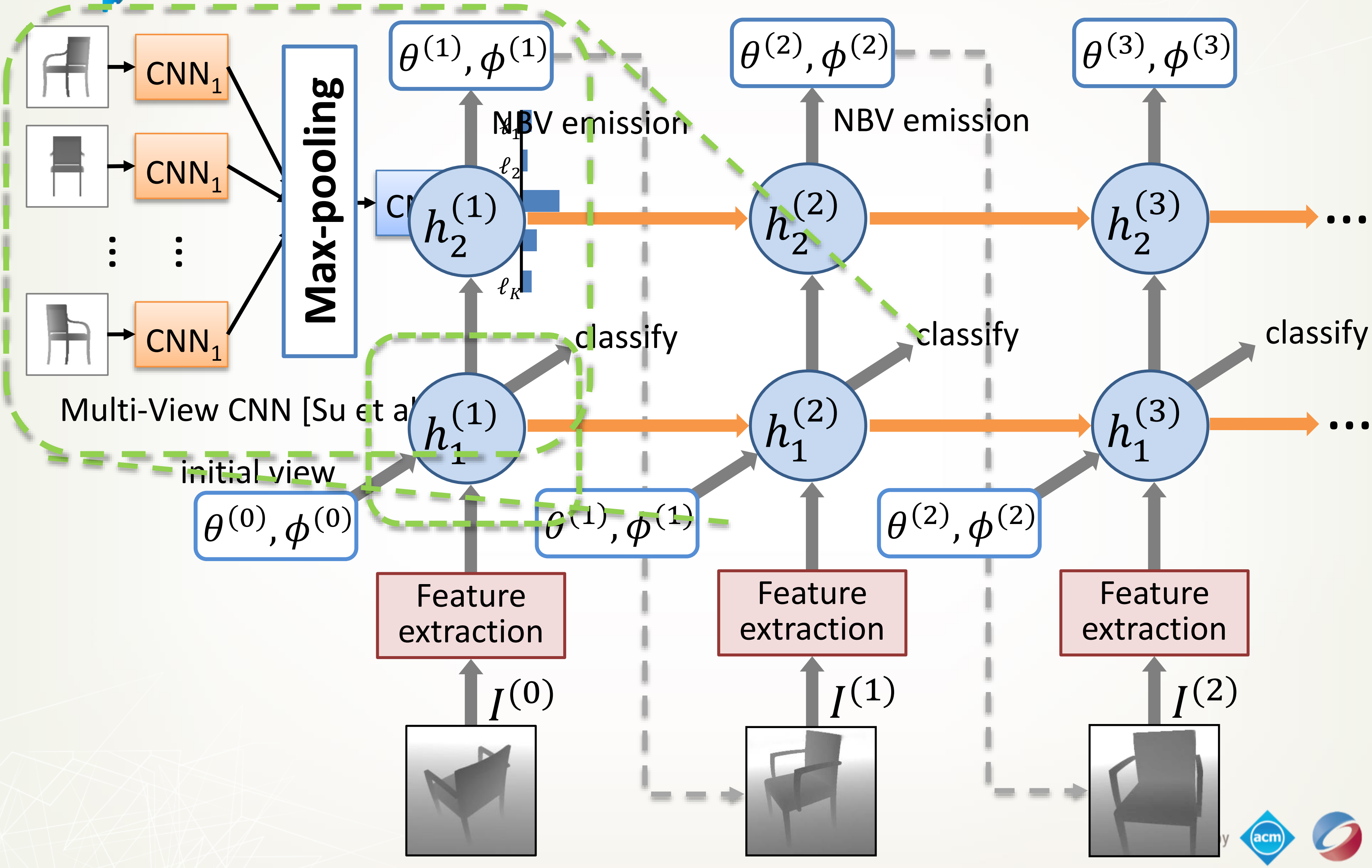
View-based observation



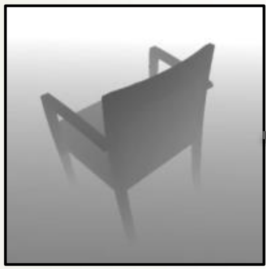
3D Recurrent Attention Model



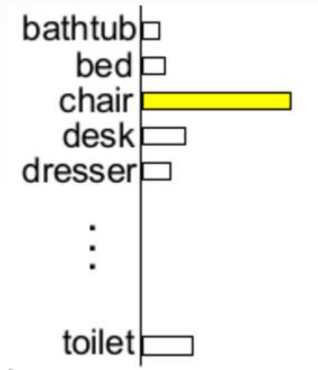
3D Recurrent Attention Model



Network training



CNN

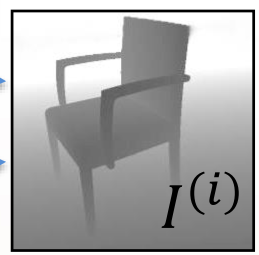


~~Back propagation~~

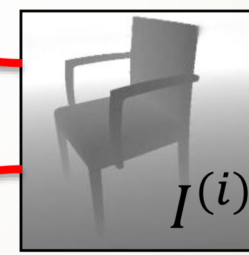
Reinforcement learning

$\theta^{(i)}, \phi^{(i)}$

rendering



$\theta^{(i)}, \phi^{(i)}$



Indifferentiable

Reinforcement learning

Stop?

agent

state

reward

action

Depth acquisition

How good the depth is?

environment

Reward

$$r_t = \underbrace{H_t(p_t, \bar{p})}_{\text{prediction accuracy}} + \underbrace{I_t(p_t, p_{t-1})}_{\text{information gain}} - \underbrace{C_t}_{\text{movement cost}}$$

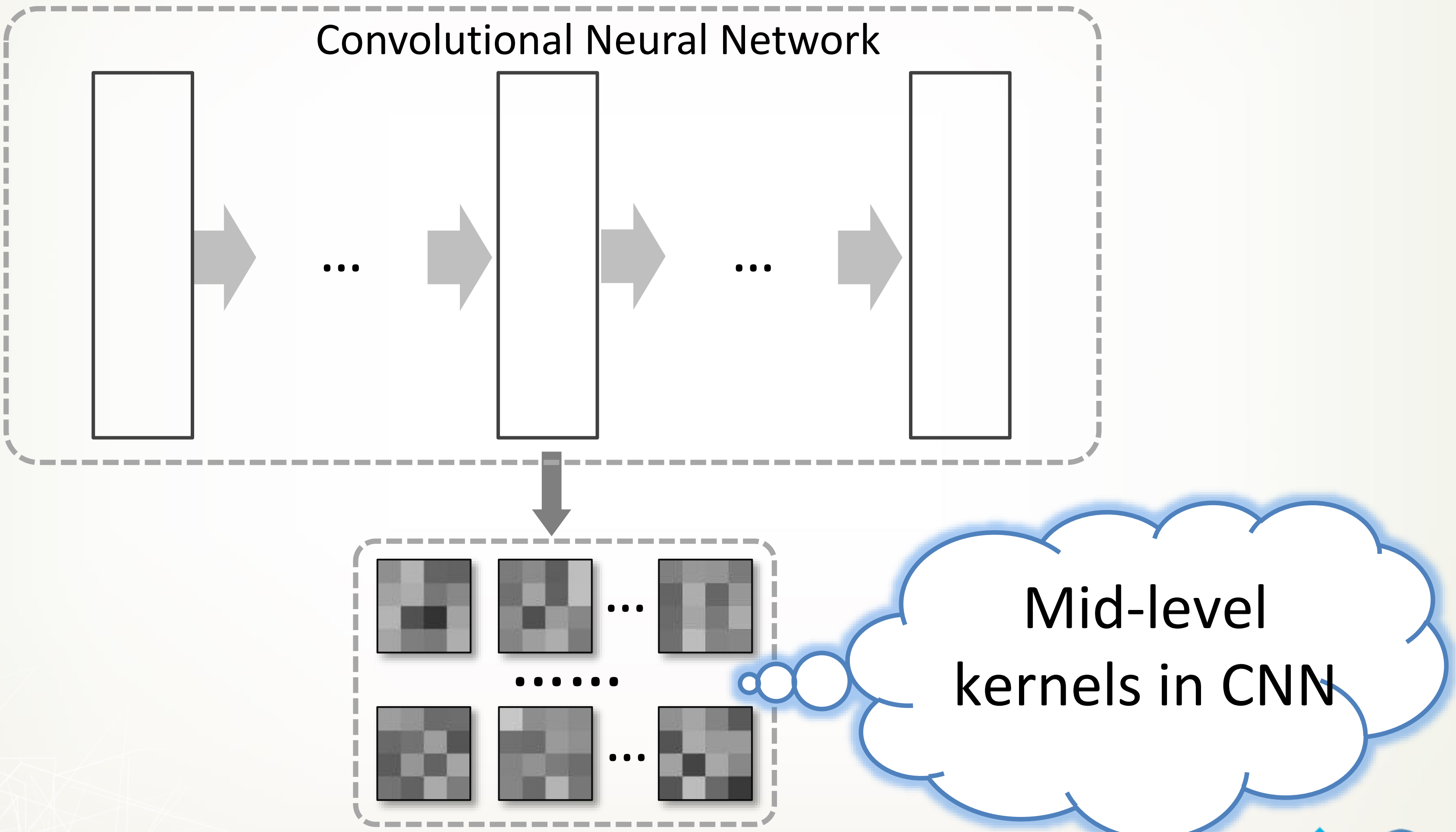
Part-level attention

occlusion

Informative parts

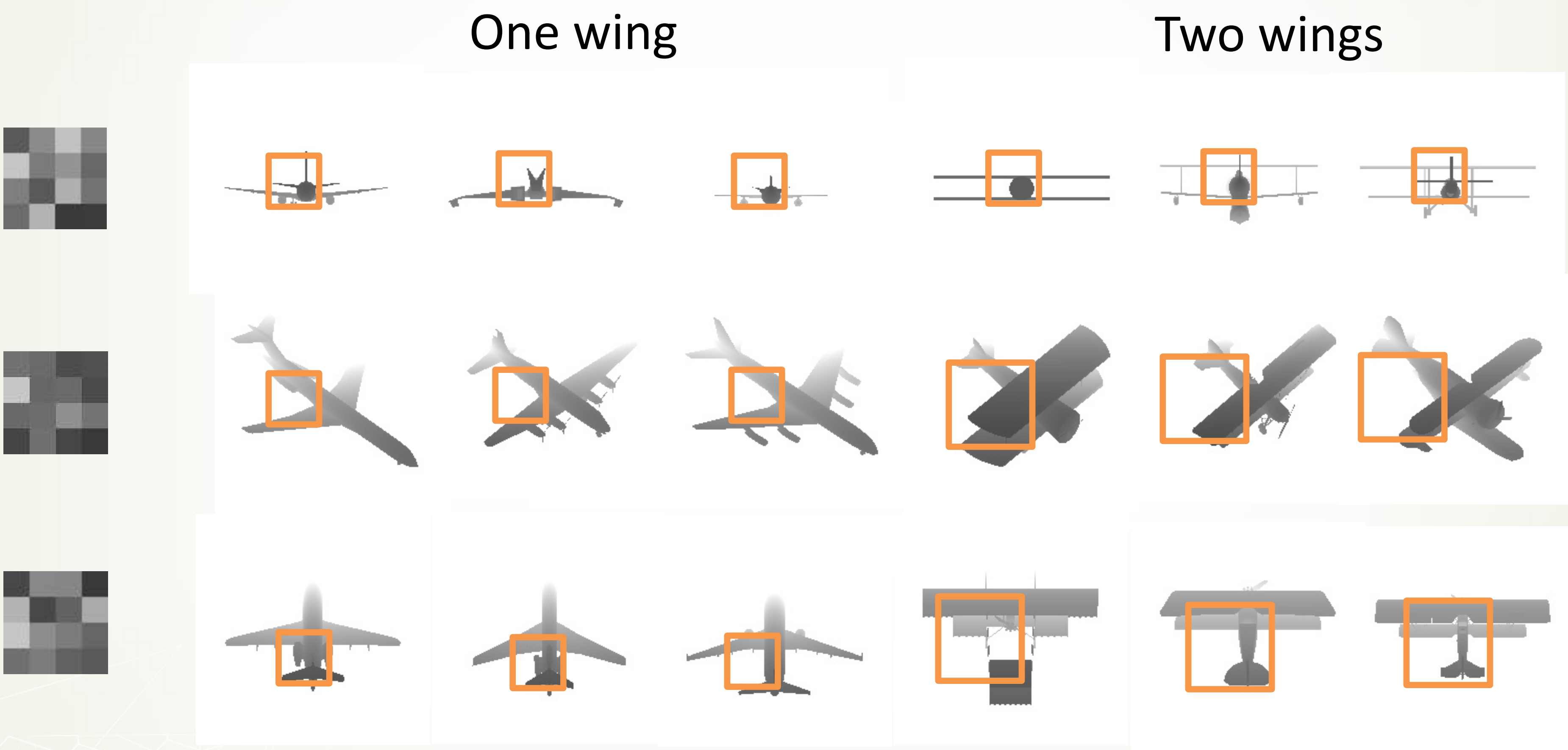
How to distinguish these two chairs?

Attention extraction



SIGGRAPH ASIA 2016 MACAO

Attention extraction



Results and evaluation

Database

SHAPENET

57,452 models
57 categories

PRINCETON MODELNET

12,311 models
40 categories

model

Render → 52 sampled views

Render with jittering → 260 sampled views

Timing

Database	MV-RNN train	MV-RNN test
ShapeNet	49 hr.	0.1 sec.
ModelNet40	22 hr.	0.1 sec.

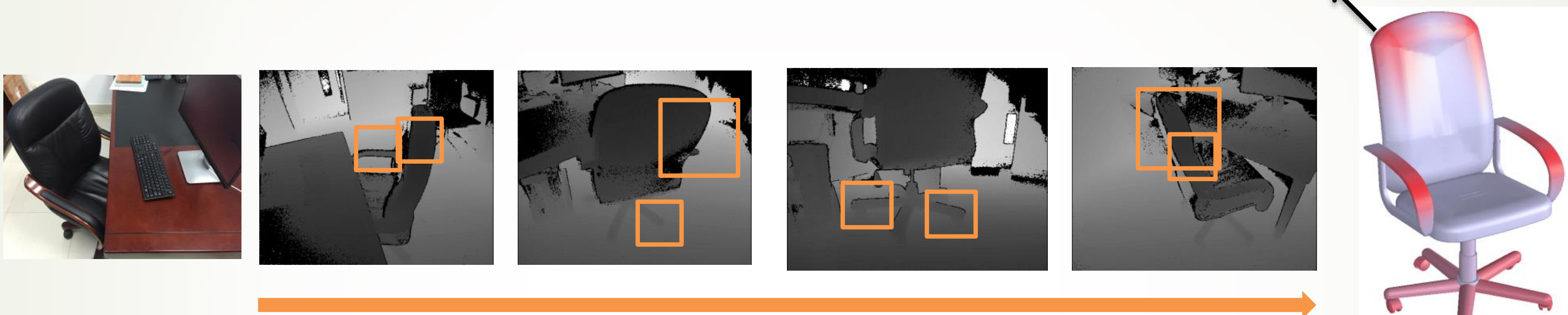
torch

A SCIENTIFIC COMPUTING FRAMEWORK FOR LUAJIT

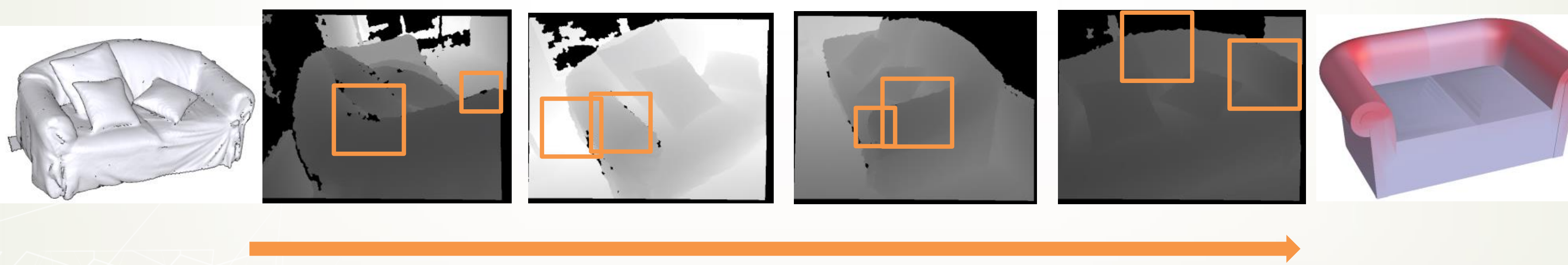
ROS

Visualization of attentions

Part-level attention



View sequence

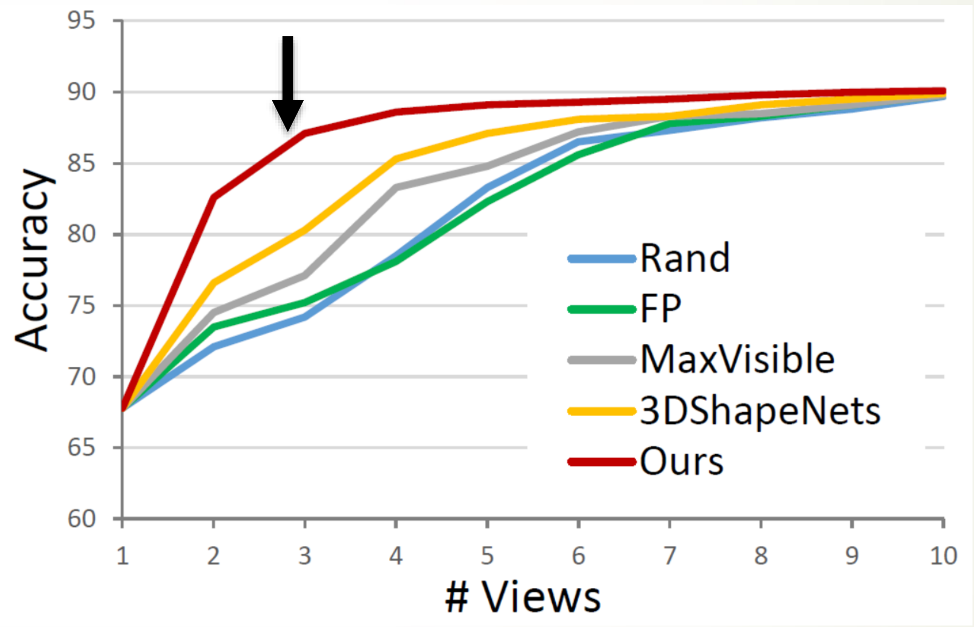


View sequence

NBV estimation

PRINCETON MODELNET

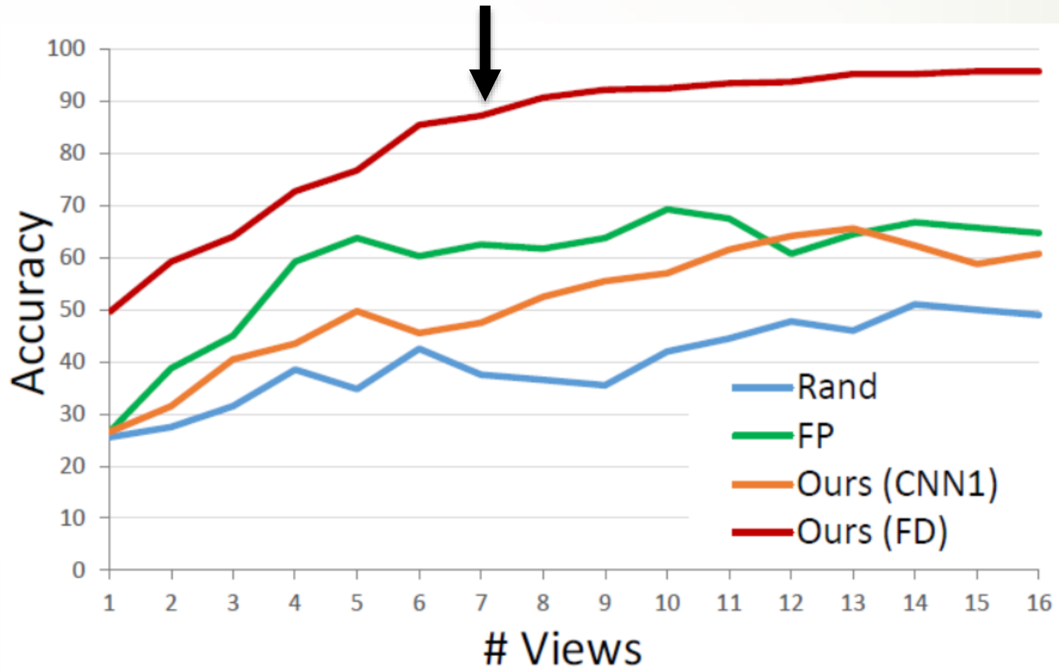
40 classes



Classification Accuracy

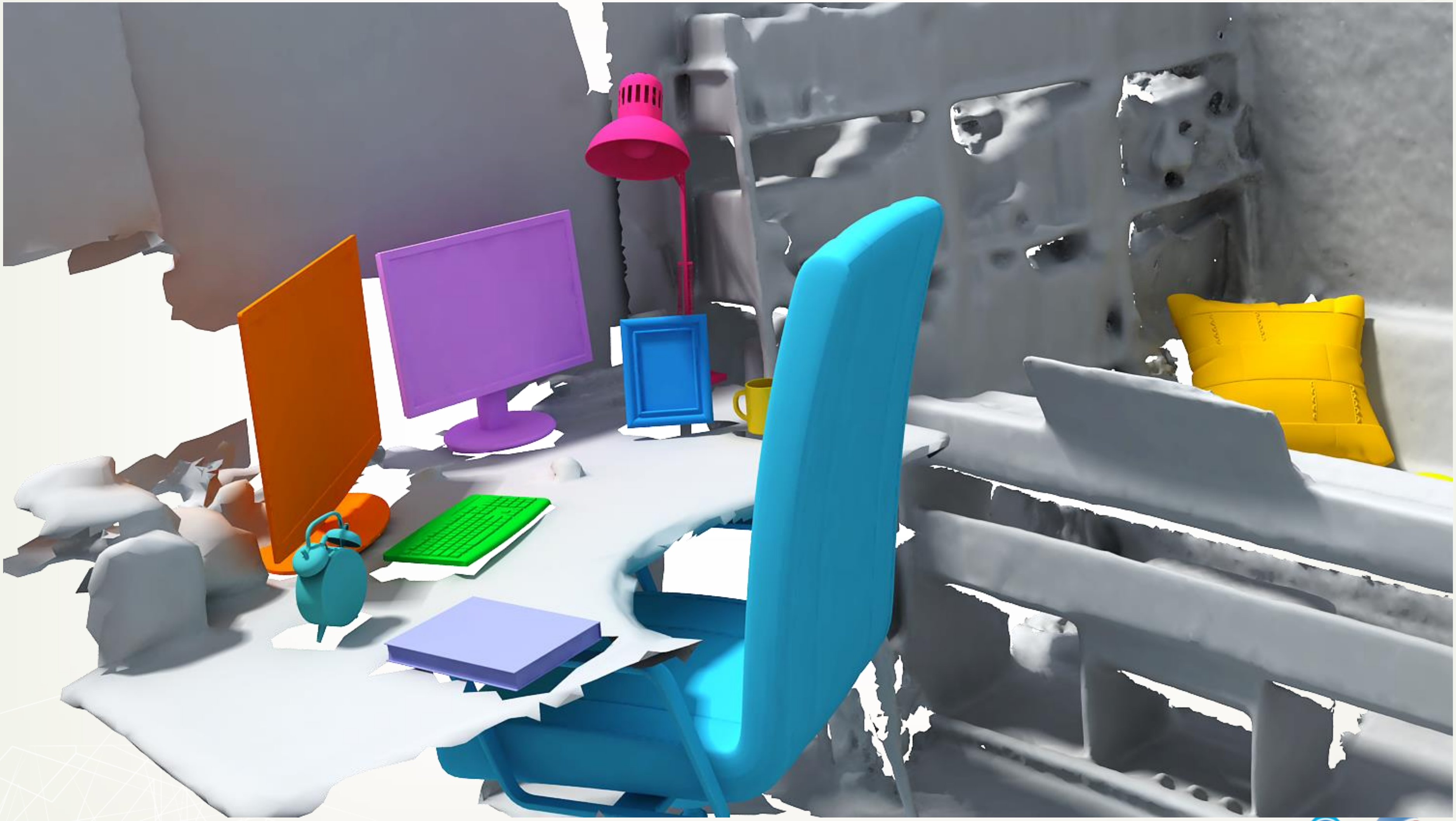
NBV estimation under occlusion

⋮



Classification Accuracy

Results on real scenes



Results on real scenes

Sponsored by

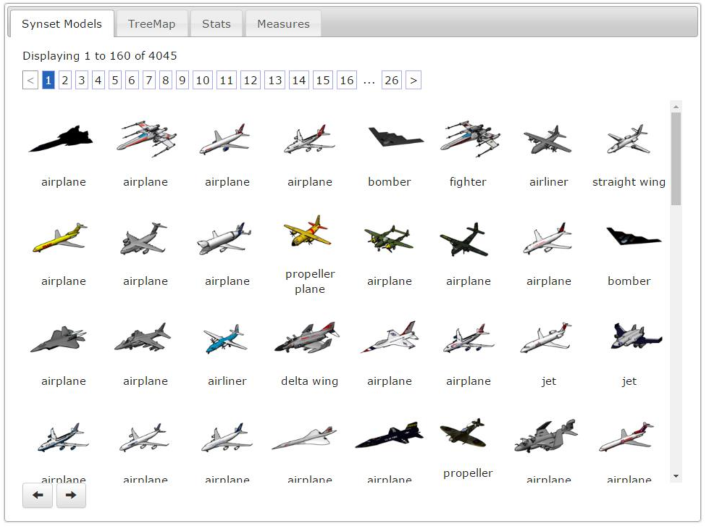
SIGGRAPH ASIA 2016 MACAO

Results on real scenes

Sponsored by acm

Limitations

- Recognizable objects
- No contextual information



Future works: Multi-modal recognition

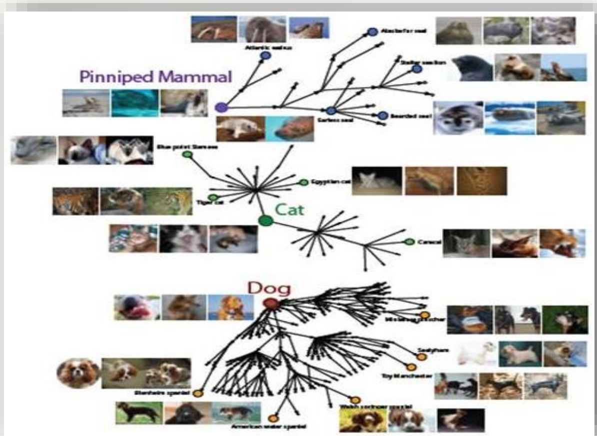
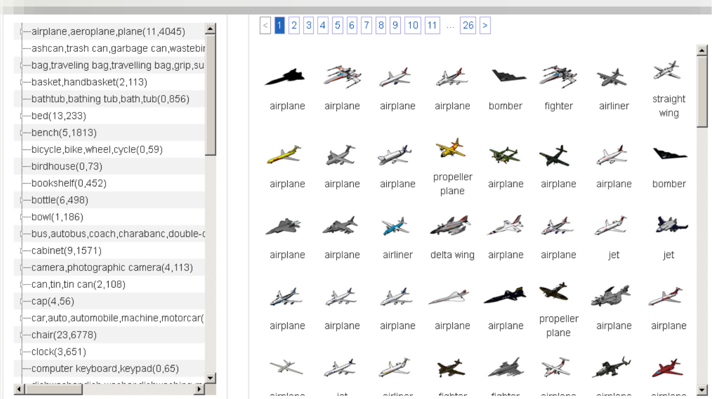
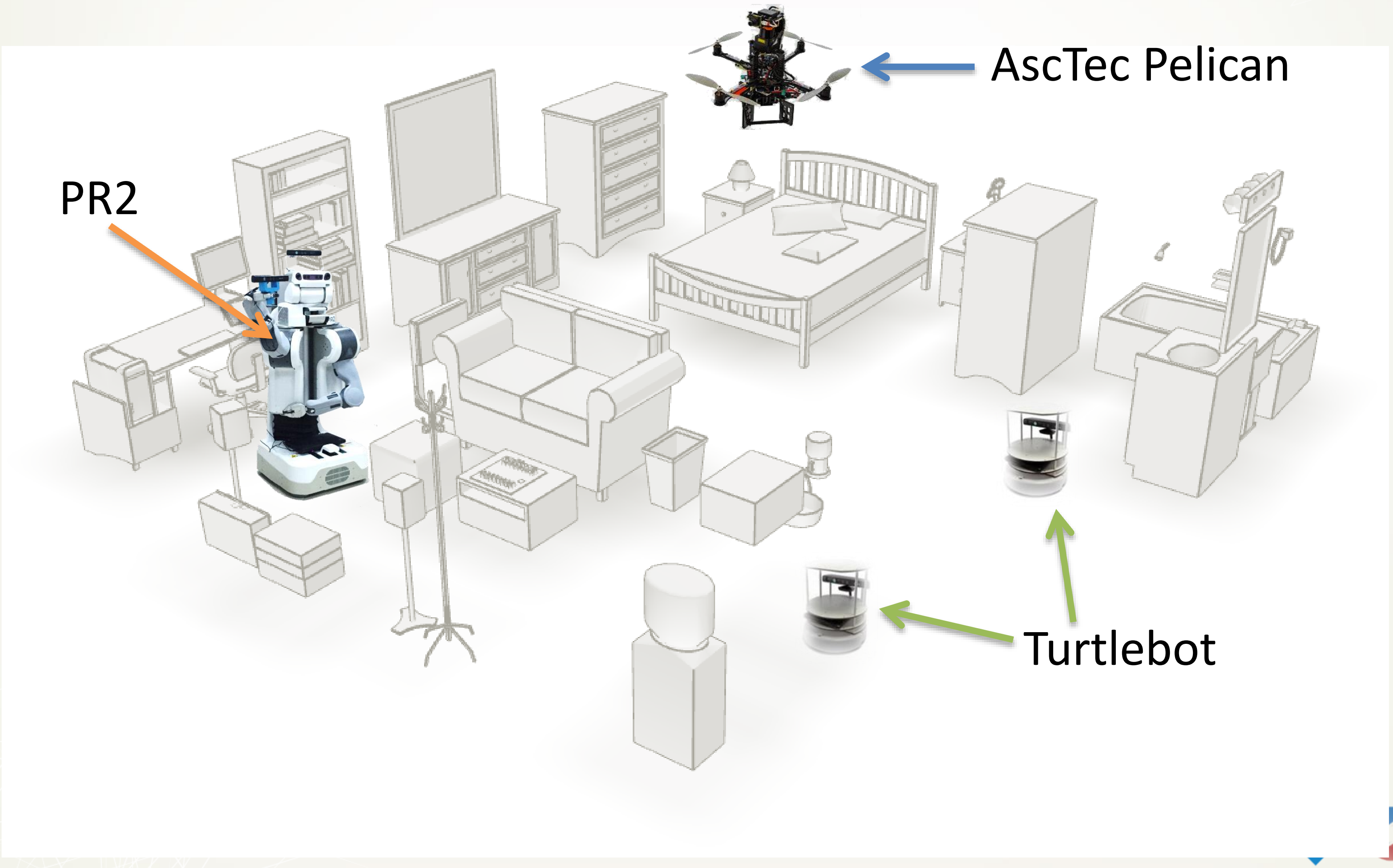


Image database



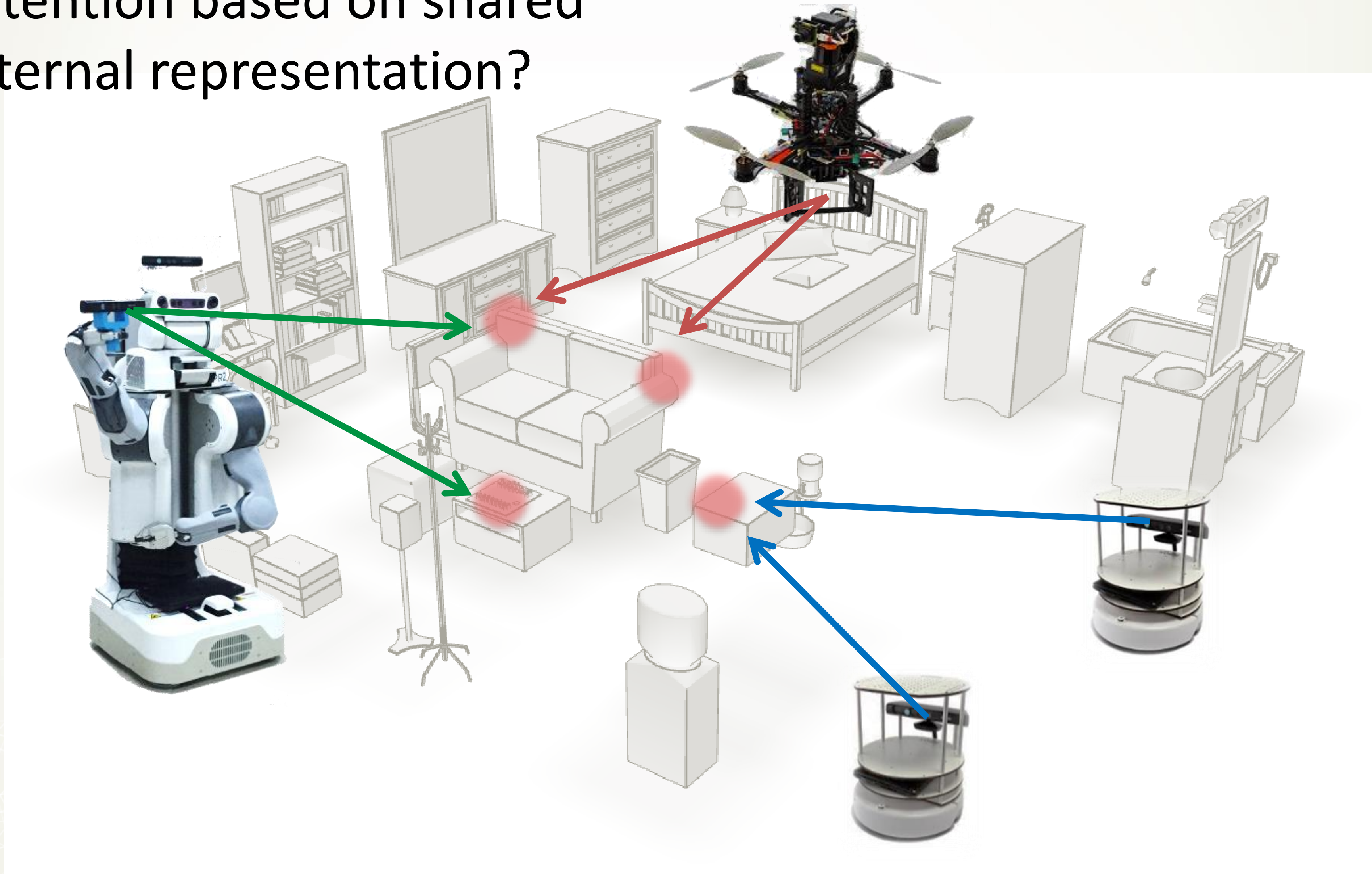
Shape database

Future: Multi-robot scene reconstruction & understanding



Future: Multi-robot attention model

Attention based on shared internal representation?



Thank you

Q & A

More details: kevinkaixu.net & yifeishi.net